Statistical Inference in Behavior Analysis: Environmental Determinants?

Nancy A. Ator
Johns Hopkins School of Medicine

Use of inferential statistics should be based on the experimental question, the nature of the design, and the nature of the data. A hallmark of single-subject designs is that such statistics should not be required to determine whether the data answer the experimental question. Yet inferential statistics are being included more often in papers that purport to present data relevant to the behavior of individual organisms. The reasons for this too often seem to be extrinsic to the experimental analysis of behavior. They include lapses in experimental design and social pressure from colleagues who are unfamiliar with single-subject research. Regardless of whether inferential statistics are used, behavior analysts need to be sophisticated about experimental design and inferential statistics. Such sophistication not only will enhance design and analysis of behavioral experiments, but also will make behavior analysts more persuasive in presenting rationales for the use or nonuse of inferential statistics to the larger scientific community.

I am by no means a sophisticated statistician and never liked math. In fact, it was a source of great relief as a graduate student to realize that my chosen enclave of research in psychology not only did not use inferential statistics but had well-founded and eloquently stated conceptual reasons for eschewing both them and the realm of "hypothesis testing" itself (Sidman, 1960; Skinner, 1950). During my tenure as an associate editor of the *Journal of the Experimental Analysis of Behavior* (*JEAB*), I found, with some dismay, that people were submitting papers that included statistical analyses to *JEAB,* the bastion of single-subject design in basic research. I had long since faced the fact that I did need to learn something about statistics and had even included an analysis of variance (ANOVA) or two in my own manuscripts; but the editorial responsibilities provided the impetus to really think about the issues of *why?, when?,* and *which?* I began looking at statistics in all manuscripts (not just *JEAB* manuscripts) in a different light and listening to colleagues talk about their approaches to data analysis from a new perspective.

As a light-hearted summary of what I have seen and heard, I present below the top 10 reasons given for using inferential statistics:

10. "My experiment used a truly randomized design."

9. "I couldn't use a truly randomized design for practical reasons, but I planned subject assignment in advance to compensate for nonequivalence of groups."

8. "I'm doing research in a clinical setting. I plan to do a time-series analysis with my single-case data, because I have limited flexibility in conducting reversals and manipulating parameters of the treatment."

7. "I'm doing single-subject research with college students because I won't need as many as a group design; but there's no way I can run all the conditions to stability or do replications and be finished by the end of the semester. So, I figure I can do an ANOVA on the group data as a back-up to the individual graphs."

6. "I thought I was doing a single-subject design with my rat study, but I wasn't able to keep up with the data

Preparation of this manuscript was supported by Grant RO1-DA04133 from the National Institute on Drug Abuse.

Reprints can be obtained from the author at the Behavioral Biology Research Center, Johns Hopkins Bayview Campus, 5510 Nathan Shock Drive, Suite 3000, Baltimore, Maryland 21224-6823 (E-mail: ator@mail.jhmi.edu).

before running tests, and it turns out the baselines were really different across rats, and I can't tell if there's anything there, and I need to get a publication out of this!"

5. "I'm an assistant professor working towards tenure. Even though I'm committed to an *experimental* analysis of behavior, I need to throw in some statistics because a senior faculty member gave me a really hard time at a departmental seminar about how I could make anything out of so few subjects."

4. "Look, my last grant application got shot down in study section because I didn't include any plan for inferential statistics. I can't afford to let that happen again."

3. "My last manuscript got shot down by a reviewer who wasn't convinced that I had a reliable effect and wanted to see some statistics."

2. "I really think these single-subject data are best suited for Journal XYZ, but I hear the new editor is biased against papers that don't include statistics."

1. "The guy down the hall got this great software package that gave me a $p < .001$, so I think I'll include it!"

These reasons can be separated easily into those that are legitimate and those that are less so. They can be categorized as ones for which the rationale for using inferential statistics is intrinsic to the nature and design of the research and those for which the rationale is extrinsic to the experimental question. In the remainder of the paper, I will discuss the issues raised in the list above. I will conclude with what seems to me to be the antidote to extrinsic determinants of the use of inferential statistics in the experimental analysis of behavior. (I must admit that some of these extrinsic reasons have affected even my own behavior over the years.)

INTRINSIC REASONS FOR USING STATISTICS

Inferential statistics are, of course, appropriate for true group designs, that is, for experiments that use random assignment of subjects to conditions or conditions to subjects, and plan to control for variability via statistical methods. Too, there are procedures for handling research situations in which subjects cannot be assigned randomly so that inferential statistics still are appropriate. Reasons 10 and 9, which describe classical group designs, are, of course, recognized as necessary conditions for use of inferential statistics (Bordens & Abbott, 1996; Rosnow & Rosenthal, 1996).

Reason 8, which refers to behavioral treatment research, also is a legitimate rationale for inferential statistics. Within the world of single-subject or single-case designs, there are appropriate statistical methods to aid evaluation of treatments in which conditions, usually in clinical settings, are not optimal for experimental control. Texts on statistics for single-case designs discuss the problems and pitfalls of such research (e.g., limitations on reversals or the ability to manipulate parameters). They set forth ways in which statistics can help researchers draw appropriate conclusions from data collected in settings in which true *experimental* analysis is not possible (Bordens & Abbott, 1996; Kazdin, 1984; Krishef, 1991).

EXTRINSIC REASONS FOR USING STATISTICS

Reasons 7 through 1 are problematic. Whether you agree with behavior-analytic colleagues who see a useful role for inferential statistics within the experimental analysis of behavior or not, these seven reasons are extrinsic to sound science.

Faulty Planning, Real Life, and the Desire to Publish

The art of experimental design, whether group or single-subject, sometimes seems to be falling by the wayside. The publication manual of the American Psychological Association used to have "Design" as the first section of "Methods"; now the term is not even in the index. Reasons 7 (research using college students) and 6 (research using rats) exemplify those situations in which experiments are not planned with an eye to the strongest possible design for the experimental question.

The single-subject design has a practical appeal over group designs because it requires fewer subjects, does not require randomization, and discourages the practice of gathering pilot data (Sidman, 1960). So, it is sometimes too easy to get started on an experiment—perhaps even in the spirit of Skinner's (1956) famous first unformalized principle of scientific practice, "When you run onto something interesting, drop everything else and study it."

The rub comes when one is faced with the labor and time commitment involved in *experimental* analysis of behavior: stability criteria, the number of reversals needed to manipulate an independent variable, equating baselines across subjects, parametric manipulations, and close monitoring for long periods of time. Then, complications occur: aging rats, equipment that malfunctions, baselines that drift, unexpected variability across subjects, and unexpected order effects. All this labor and these complications occur in the context of academic deadlines, funding deadlines, promotion and tenure reviews, and other realities of life.

It is no wonder that there is great appeal in taking what data have been collected, putting them through a few statistical manipulations, and seeing whether there is "anything there." Some of these manipulations *can* turn out pretty well, and, regardless of the design, well-collected, interesting, and clear data should be published. Sometimes though, the result is neither fish nor fowl—a hybrid single-subject/group design with the best of neither: few subjects, great variability, little replication, few parametrics, mean data that are "significant" but represent few if any of the subjects, analyzed with statistical procedures that are arguably inappropriate for the data. Sometimes, this approach has been encouraged by another class of environmental determinants: the academic social environment.

The Devil Made Me Do It

Reasons 5 through 2 describe socially mediated contingencies that support inclusion of statistical analyses. As behavior analysts make their way in the academic world, it is a fact of life that at one time or another, our research presentations are questioned for their lack of *p* values. Adding statistics then becomes an avoidance response that heads off criticism from colleagues who refuse to take seriously any result not accompanied by "$p <$ .05."

When behavior analysts do not understand design and statistical methods well enough, we become unduly subject to the preconceived notions of reviewers, editors, colleagues, and department chairs, who are not trained to appreciate steady-state research. By unduly subject, I mean that we cannot stand up for single-subject designs in a credible way. Many of the very people (reviewers and editors) who call for more statistics do not understand them either, and the contingency seems to be placed on having a *p* value. To the extent this is true, many of the statistics reported in psychological journals seem to represent rule-governed behavior run amok! To be able to argue effectively against such rule-governed behavior, behavior analysts who believe inferential statistics to be inappropriate for their data must gain a thorough understanding of how such statistics should be used and what they can and cannot do (Branch, 1999; Perone, 1999).

Beware the Gold Star

Reason 1, the great *p* value provided by statistical analysis software, while the most facetious, may be the most insidious. Because good steady-state research with lots of control of environmental variables (not to mention the "salutary" influence of autocorrelation) produces results with clear differences in effects, it turns out to be remarkably easy to find significant *p* values even with few subjects. With easy-to-use statistical software, who can fail to enjoy the immediate reinforcement of plugging data into a spreadsheet and, in the wink of an eye and the click of a mouse, seeing "$p = .0001$." Like pasting a gold star on your data! The emergence of symposia, journal articles, and a task force questioning the unwarranted "significance" of *p* values, however, should suggest caution (cf. Harris, 1997; Hopkins, Cole, & Mason, 1998; Loftus, 1996). The wind may be changing.

CONCLUSION

Rather than make an appeal for or against a role for statistics in behavior analysis, I want to make an appeal for thoughtfulness in experimental design and for being more sophisticated in our knowledge of statistics. Use of inferential statistics should be based on the experimental question, the nature of the design, and the nature of the data.

A hallmark of single-subject designs in the experimental analysis of behavior is that such statistics should not be required to determine the reality of the effect of an independent variable. Although there are situations in which inferential statistics can be useful adjuncts to visual inspection of the data in single-subject designs (Fisch, 1998; Krishef, 1991), the use of statistics for reasons that are extrinsic to the experimental design should be minimized. Sometimes use of statistics is the result of poor planning or execution of an experiment: a quasi-single-subject design. Although efforts to salvage data that have been carefully collected are defensible, the statistics as used often are not.

Behavioral science can only be strengthened by a decline in the kind of social variables suggested in Reasons 7 through 1 as primary determinants of using a *p* value. To counteract these social influences, students of the experimental analysis of behavior should be taught statistics as thoroughly as possible, and the rest of us should brush up. Behavior analysts need to be sufficiently sophisticated about the experimental designs they use to be able to argue persuasively for the most appropriate analysis of the data, given the design they have chosen, and to resist using inferential statistics where inappropriate. In particular, we should be proactive in setting forth our choices of experimental design and our rationales for concluding that effects did or did not occur. This should be true of our manuscripts and, most especially, of our research grant proposals. Finally, we should be able to review manuscripts well enough to understand whether the statistics included are appropriate and appropriately described. Benefits to the field include more solidly based conclusions in the literature and perhaps greater respect for single-subject designs from colleagues who now dismiss them.

REFERENCES

Bordens, K. S., & Abbott, B. B. (1996). *Research design and methods: A process approach* (3rd ed.). Mountain View, CA: Mayfield.

Branch, M. N. (1999). Statistical inference in behavior analysis: Some things significance testing does and does not do. *The Behavior Analyst, 22,* 87–92.

Fisch, G. S. (1998). Visual inspection of data revisited: Do the eyes still have it? *The Behavior Analyst, 21,* 111–123.

Harris, R. J. (Ed.). (1997). Special section: Ban the significance test? *Psychological Science, 8,* 1–20.

Hopkins, B. L., Cole, B. L., & Mason, T. L. (1998). A critique of the usefulness of inferential statistics in applied behavior analysis. *The Behavior Analyst, 21,* 125–137.

Kazdin, A. E. (1984). Statistical analyses for single-case experimental designs. In D. H. Barlow & M. Hersen (Eds.), *Single case experimental designs: Strategies for studying behavior change* (2nd ed., pp. 265–316). New York: Pergamon Press.

Krishef, C. H. (1991). *Fundamental approaches to single subject design and analysis.* Malabar, FL: Krieger.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5,* 161–171.

Perone, M. (1999). Statistical inference in behavior analysis: Experimental control is better. *The Behavior Analyst, 22,* 109–116.

Rosnow, R. L., & Rosenthal, R. (1996). *Beginning behavioral research: A conceptual primer* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology.* New York: Basic Books.

Skinner, B. F. (1950). Are theories of learning necessary? *Psychological Review, 57,* 193–216.

Skinner, B. F. (1956). A case history in scientific method. *American Psychologist, 11,* 221–233.